AutoAnnotator: An Agentic AI Framework for Multimodal Data Annotation

Mrs. Farhat A. Patel[1], Aditi Ramdas Dhum[2], Devyani Dadaram Badgujar[3], Sakshi Kailas Jadhav[4], Neha Aniruddha Joshi[5]

[1]*Internal Guide, Department of Computer Engineering, Guru Gobind Singh College of Engineering and Research Centre, Nashik*

[2,3,4,5]*Students, Department of Computer Engineering, Guru Gobind Singh College of Engineering and Research Centre, Nashik*

***Abstract*—This paper reviews 20 selected publications on automated and AI-assisted data annotation, covering images, audio, documents and 3D data, alongside research on agentic AI. The reviewed studies are compared thematically by modality, annotation task, model and automation approach, verification mechanism, human involvement, evaluation practice and reported limitations. Across the reviewed corpus, pre-trained and foundation models support tasks such as pseudo-labelling, zero-shot prediction and caption generation. However, verification remains task-specific, using approaches such as confidence-based human routing, multi-model voting, question-answering self-checks and comparison with reference annotations. The corpus does not contain a demonstrated lightweight workflow that coordinates annotation and verification across image, video and audio, or an annotation system explicitly orchestrated by an agent/planner. Based on these observations, the paper discusses the proposed AutoAnnotator framework, which combines modality detection, an Agent/Planner, suitable pre-trained models for image annotation, video detection and tracking, and audio speech-to-text, with confidence-based verification and structured JSON/CSV export. The framework is conceptual: it has not been implemented or experimentally evaluated in this paper, and no performance results are claimed.**

***Index Terms*—agentic AI, automated data annotation, confidence-based verification, human-in-the-loop, multimodal annotation, vision-language models**

I. INTRODUCTION

Large labelled datasets are a prerequisite for supervised learning, and annotation quality directly affects model quality [1], [17]. Manual annotation is labour-intensive, and its quality depends on annotator expertise, guidelines and task design [1]. The cost is especially visible in 3D and specialised domains, where dense labelling is hard to scale [10], [11], [15], [17].

Two trends address this problem. The first is model-based automation, in which pre-trained classifiers, vision-language models (VLMs), large language models (LLMs) and multimodal LLMs generate labels, masks, captions or pseudo-boxes [1]– [6], [12]–[16]. The second is agentic AI, in which LLM-based systems plan, use tools and act through iterative feedback loops [7], [8]. The reviewed literature contains a range of single-task annotation methods and broader agent research, but their integration into a unified cross-modal annotation workflow remains an open question within the corpus examined here.

The AutoAnnotator project synopsis proposes a low-cost, agent-based workflow that identifies the input type (image, video or audio), applies a suitable pre-trained model for the selected annotation task, checks outputs using confidence values and simple consistency rules, and exports results in JSON or CSV. The proposal builds on existing pre-trained models rather than training new models from scratch.

Contributions. This paper (i) synthesises 20 selected publications using a common set of analytical dimensions, (ii) identifies research gaps within the reviewed corpus, and (iii) discusses the proposed AutoAnnotator framework in relation to those findings. The review findings and the proposed framework are kept distinct.

The remainder of the paper is organised as follows. Section II describes the review scope and method.

Section III synthesises the literature, followed by comparative analysis in Section IV and verification strategies in Section V. Section VI identifies gaps within the reviewed corpus. Section VII presents the proposed AutoAnnotator framework, and Sections VIII–X discuss implications, future directions and conclusions.

II. REVIEW SCOPE AND METHOD

Evidence base. This paper examines a defined corpus of 20 publications selected for relevance to automated annotation, verification and agentic AI, together with the AutoAnnotator project synopsis. The review is not intended to provide exhaustive coverage of the field. The analysis is based on information reported in the publications; the reported results were not independently reproduced.

Corpus. The publications span 2017–2026 and include IEEE, ACM/IEEE and CVF conference and journal papers, as well as three 2026 arXiv preprints [4]–[6]. The corpus is heterogeneous: some papers directly address annotation [1]–[6], [9]–[11], [17], while others contribute annotation-free learning [12]–[16], foundational 3D or segmentation methods [18]–[20], or agentic-AI concepts [7], [8].

Analytical dimensions. The publications were examined in terms of modality, annotation task, models and techniques, automation level, agentic elements, confidence and verification, human involvement, datasets and evaluation, reported limitations, and relevance to AutoAnnotator. These dimensions guide the comparison and discussion throughout the paper.

Boundaries. Most 3D and point-cloud papers [10]–[15], [17]–[19] fall outside AutoAnnotator's proposed image, video and audio scope, but inform the discussion of automation and verification. The reviewed corpus provides limited direct evidence on speech-to-text annotation and video object tracking.

III. LITERATURE SYNTHESIS

A. AI-assisted annotation of images

The image-related studies cover expert-assisted satellite annotation, comparative image classifiers, zero-shot livestock labelling and biomedical segmentation. Beck et al. [1] report that certainty-based routing can reduce manual cost while maintaining high agreement, although calibration and prompt sensitivity remain concerns. Jain et al. [2] compare convolutional neural network (CNN) feature extractors and CLIP (Contrastive Language–Image Pre-training) features with standard classifiers, illustrating the potential of transfer-based image annotation. Su et al. [3] combine multiple vision-language models, segmentation and agreement checks for zero-shot labelling. Zhang and Sun [20] present a U-Net-derived biomedical segmentation method. Taken together, these studies suggest that automated image labelling can reduce manual effort, but reliability depends on the task, model and verification method. Fig. 1 illustrates a typical form of image-annotation output.

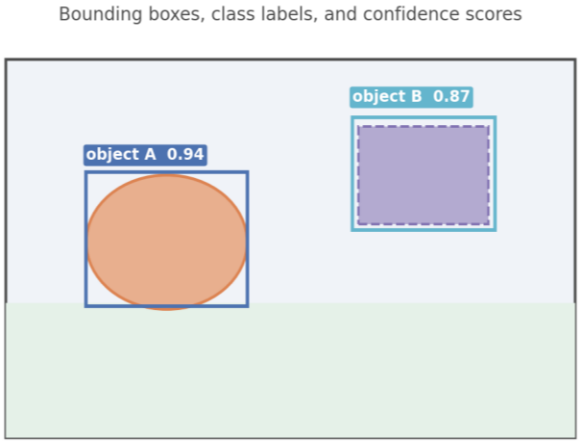

Fig. 1. Illustrative example of image annotation output: bounding boxes, class labels, and confidence scores for detected objects (schematic, not drawn from a specific paper).

B. AI-assisted annotation of audio

The audio-related studies address caption generation and structure detection rather than label assignment alone. Lahrichi et al. [4] generate diverse, fine-grained audio captions by prompting a multimodal large language model, have them rated in a human evaluation, and use the resulting dataset to train CLAP models for audio retrieval. Alex et al. [5] detect chapter boundaries in media audio by post-training large audio-language models with group relative policy optimisation (GRPO) against creator-authored chapter annotations. Ma et al. [6] add a verification loop to long-paragraph audio captioning, in which

question-answering self-checks test whether the caption is supported by the audio. Taken together, these studies show that audio annotation can be automated with foundation models and that verification is typically built around the generated text. The reviewed audio studies do not directly evaluate plain speech-to-text transcription, which is the audio task proposed for AutoAnnotator.

C. 3D, document and annotation-free approaches

Several papers fall outside AutoAnnotator's image, video and audio scope but show how annotation effort can be reduced elsewhere. Semi-automatic and assistive tools support 3D point-cloud and microscopy annotation, where humans still confirm or correct labels [10], [11], and a content-reuse annotator addresses text, images and mathematics in documents [9]. Annotation-free methods transfer knowledge from 2D vision-language and segmentation models to 3D or interaction tasks, avoiding manual 3D labels altogether [12]– [16]; for example, [13] and [14] train open-vocabulary 3D detectors without 3D annotation. ScanNet [17] illustrates the scale of human effort behind richly annotated 3D data, while PointNet and MVPointNet [18], [19] are supervised models that need such labels. These works inform the discussion of automation and verification, but they do not provide direct evidence for AutoAnnotator's proposed modalities.

D. Agentic AI foundations

Ferrag et al. [7] describe agents as systems that add goals, memory, tool use, closed-loop interaction and adaptive control around reasoning models. Ishida et al. [8] connect classical agent architectures with modern AI-agent implementations and discuss tool use and orchestration. These works provide conceptual background for the Agent/Planner in AutoAnnotator. Within the reviewed corpus, however, the annotation-focused papers do not demonstrate an agent/planner coordinating annotation across image, video and audio.

IV. COMPARATIVE ANALYSIS

Table I compares the modalities, tasks and techniques of selected publications, while Table II compares their automation levels, verification or confidence mechanisms, and human roles. The tables complement the discussion of the 20-paper corpus by highlighting technical approaches and differences relevant to the proposed AutoAnnotator framework.

Table I. Modalities, tasks and techniques.

Ref.	Modality	Annotation / core task	Models and techniques
[1]	Images (satellite)	Satellite image annotation: label quality and cost	ChatGPT (multimodal LLM) vs human annotators; certainty-based routing
[2]	Images	Product-category labelling	ResNet50, Xception, InceptionV3, CLIP-ViT-B/32; LR, XGB, KNN, RF, SVM
[3]	Images	Zero-shot instance masks	Qwen2.5-VL, Molmo, Gemini 2.5 Pro, SAM2, IoU fusion, VLM voting
[4]	Audio	Diverse audio captions	Qwen3-Omni prompting; CLAP
[5]	Audio	Chapter-boundary detection	AF3-Think-8B, Step-Audio-R1, Gemini 2.5 Pro; SFT + GRPO
[6]	Audio	Long-paragraph detailed audio captioning	Self-check captioning with question-answering verification
[7]	Multimodal (driving, UAV)	Survey: reasoning, agents, datasets, security	LLM reasoning and agent frameworks
[8]	Concept (AI agents)	Agent architectures, tool use, orchestration	Classical agent architectures; LLM-based agent implementations

Ref.	Modality	Annotation / core task	Models and techniques
[9]	Documents (text, images, math)	Content-reuse annotation	Annotation tool (TEIMMA)
[10]	3D (multi-modal streams)	Semi-automatic 3D annotation	Web-based annotation toolbox (3D BAT)
[11]	3D (microscopy)	Ground-truth annotation for nuclei segmentation	Annotation tool and 3D dataset generator
[12]	3D scenes	Open-vocabulary 3D scene understanding	OpenScene: CLIP-based open-vocabulary features
[13]	Point clouds	Open-vocabulary 3D detection without 3D annotation	Annotation-free training using 2D vision-language knowledge
[14]	Multi-view RGB(D)	Open-vocabulary 3D detection	SAM segments, graph-clustered pseudo-boxes, CLIP alignment
[15]	3D scenes	Open-vocabulary 3D scene graph generation	CLIP-driven; cross-modality contrastive learning
[16]	Images	Human–object interaction detection	Natural-language supervision (CLIP-based transfer)
[17]	3D indoor scans	Richly annotated reconstruction dataset (ScanNet)	Dataset built with human annotation
[18]	3D point clouds	3D classification and segmentation	PointNet (deep learning on point sets)
[19]	3D point clouds	3D object recognition	MVPointNet (multi-view network)
[20]	Biomedical images	Image segmentation	Fully convolutional network (U-Net-derived)

Table II. Automation, verification/confidence and human role.

Ref.	Automation level	Verification / confidence mechanism	Human role
[1]	Semi-automatic (LLM with routing)	Model certainty; comparison with expert/reference annotations	Humans handle low-certainty cases
[2]	Automatic classification	Held-out accuracy, precision, F1	None (future: self-correcting mechanisms)
[3]	Automatic zero-shot	Multi-model voting; redundancy removal; agreement checks	None reported
[4]	Automatic generation	Structured prompting; human evaluation; retrieval and zero-shot classification with trained CLAP models	Human evaluation
[5]	Trained model	Sanitised perception log; rule-based reward	Creator-authored boundaries as supervision
[6]	Automatic generation	Question-answering self-check	None reported
[7]	Survey (conceptual)	Not applicable; discusses hallucination and security risks	Not applicable

Ref.	Automation level	Verification / confidence mechanism	Human role
[8]	Conceptual	Not applicable	Not applicable
[9]	Tool-assisted	Tool-specific; no common output schema	User-operated tool
[10]	Semi-automatic	Comparison with reference annotations	Human annotator in the loop
[11]	Assistive	TP/FP/FN vs ground truth	Human clicks labels
[12]–[16]	Annotation-free training	Benchmarks where ground truth exists; qualitative elsewhere	None
[17]	Dataset construction	Human or crowd redundancy	Human annotators
[18]–[20]	Supervised models	Task metrics (accuracy, Dice, error)	Labels required for training

Abbreviations used in Tables I and II: CLAP, contrastive language–audio pretraining; CLIP, contrastive language–image pre-training; GRPO, group relative policy optimisation; IoU, intersection over union; KNN, k-nearest neighbours; LR, logistic regression; RF, random forest; RGB(D), colour (and depth) images; SAM, Segment Anything Model; SFT, supervised fine-tuning; SVM, support vector machine; TP/FP/FN, true positives, false positives and false negatives; UAV, unmanned aerial vehicle; VLM, vision-language model; XGB, XGBoost.

Coverage is uneven within the reviewed corpus: 3D and point-cloud work are prominent, image and audio studies are also represented, and direct evidence on video annotation is comparatively limited. This distribution motivates further investigation of workflows that coordinate tasks across modalities; it should not be interpreted as a measurement of the entire research field.

The corpus includes classical CNN/classifier pipelines, foundation and vision-language models, and multimodal or audio-language systems. Several studies use pre-trained or foundation models; AutoAnnotator similarly proposes selecting suitable pre-trained models for each task. Recurring techniques include pseudo-label transfer, multi-model complementarity, structured captioning, open-vocabulary recognition and human-assisted annotation. These techniques inform the proposed design space but are not claimed to have been implemented in AutoAnnotator.

Agentic methods are discussed in [7] and [8], whereas the annotation-focused papers in this corpus primarily describe pipelines, models or tools. Some verification loops share features with agentic workflows, but the reviewed corpus does not demonstrate an agent/planner controlling annotation across multiple modalities.

Evaluation practices vary and include agreement, mean average precision (mAP), retrieval recall, F1 score, question-answering accuracy, intersection over union (IoU) and qualitative assessment. Cost, processing time and manual-effort reduction are reported less consistently. These measures are therefore appropriate candidates for a future evaluation of AutoAnnotator, rather than results that can be inferred from the current proposal.

V. VERIFICATION AND CONFIDENCE

The reviewed publications use several verification mechanisms: model certainty with human routing [1], multi-model voting and redundancy removal [3], question-answering self-checks [6], rule-based supervision [5], comparison with expert or reference annotations [1], [10], [11], and human or crowd redundancy [1], [17]. These mechanisms are complementary and task-specific. The evidence supports treating verification as a distinct stage rather than relying on a single confidence value. AutoAnnotator therefore proposes confidence-based checks and simple consistency rules; advanced voting, question-answering checks and calibrated human-routing policies remain possible future extensions.

VI. RESEARCH GAPS

The following gaps are identified within the reviewed corpus only; they are not claims that no such work exists anywhere in the wider literature.

- *G1 No unified multimodal workflow.* Within this corpus, no reviewed paper presents one workflow that identifies the modality, selects tools, annotates

and verifies across image, video and audio. The reviewed contributions instead focus on particular modalities, tasks or domains [1]–[6], [9]–[20].

- *G2 Agentic orchestration is not demonstrated for annotation.* Agentic systems are discussed in [7] and [8], but no paper in the reviewed corpus demonstrates planning and tool selection to coordinate multimodal annotation.

- *G3 Verification is modality-specific.* Voting [3], question-answering self-checks [6] and certainty-based routing [1] are reported as separate approaches; transfer of a common verification strategy across modalities is not established by this corpus.

- *G4 Confidence reliability remains unresolved.* The reviewed work identifies calibration concerns [1], while hallucination risks are relevant to LLM-based annotation and agent components [1], [4], [6], [7]. The corpus does not establish a shared confidence-calibration approach across modalities.

- *G5 Human-in-the-loop practices vary.* Human review is absent from some automated approaches and integrated into others. Within the reviewed corpus, [1] provides a quantified certainty-based routing policy; a common human-review policy across modalities is not established.

- *G6 Limited direct evidence for video tracking and speech transcription.* The reviewed corpus contains little direct evaluation of video object tracking or speech-to-text annotation, leaving these components insufficiently supported by the current review evidence.

- *G7 Evaluation is heterogeneous.* Metrics differ across modalities and tasks, while cost, time and manual effort are reported less consistently. Direct numerical comparisons across the reviewed studies are therefore limited.

- *G8 Resource requirements may constrain deployment.* Some systems use multiple large models [3], specialised training [5], [6] or substantial computing resources [14]. The reviewed corpus provides limited comparative evidence on lightweight alternatives, although [7] discusses compact on-device models.

- *G9 No common output schema is established in this corpus.* Annotation formats are tool-specific in the reviewed examples [9], [10]. Schema-based tool interfaces are discussed in [8], but a shared annotation-output schema across modalities is not demonstrated.

- *G10 Results are domain-specific.* The reviewed studies address particular domains, including satellite imagery [1], fashion [2], livestock [3] and biomedical data [20]. Their results may not generalise directly to other domains without further evaluation.

VII. PROPOSED AUTOANNOTATOR FRAMEWORK

Status. This section describes the *proposed* system as defined in the project synopsis, with pre-trained models used for the annotation tasks. It has not been implemented or evaluated here, and no accuracy, speed or effort-reduction figures are claimed.

A. Aim, scope and design principle

AutoAnnotator aims to be a simple, low-cost agentic framework for automated annotation of image, video and audio data using suitable pre-trained models. Its objectives are to study existing approaches, identify gaps, design an agent-based framework, explore generation, verification and management of annotations, and assess accuracy, processing efficiency and manual-effort reduction.

The stated scope covers image object detection and basic classification, video object detection on selected frames with tracking across frames, and audio speech-to-text. Annotations are displayed and stored as JSON or CSV.

Testing will use small or medium public datasets and user-provided samples; specific datasets will be selected during implementation. Fig. 2 illustrates the frame-wise detection with identity preserved across frames that is scoped for video.

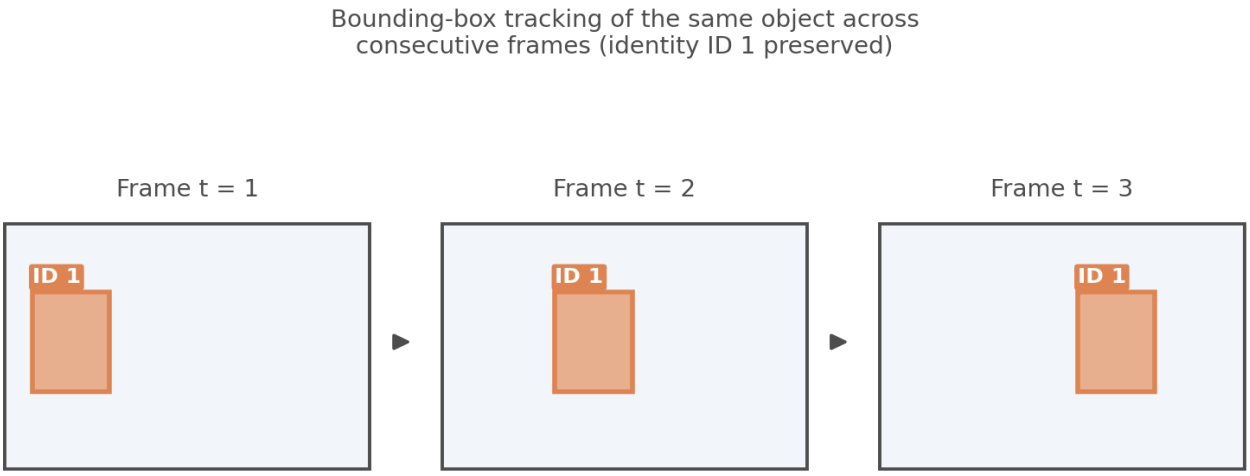

Fig. 2. Conceptual illustration of video annotation via frame-wise object detection with an identity (ID) preserved across consecutive frames, as scoped for the proposed framework.

B. Architecture and workflow

The synopsis defines ten modules: User Input, Input Detection, Preprocessing, Agent/Planner, Image Annotation, Video Annotation, Audio Annotation, Verification, Storage and Export, and Visualization. The overall architecture is shown in Fig. 3 and its components are listed in Table III; Fig. 4 illustrates the time-aligned transcript output of the speech-to-text step.

The workflow has nine steps:

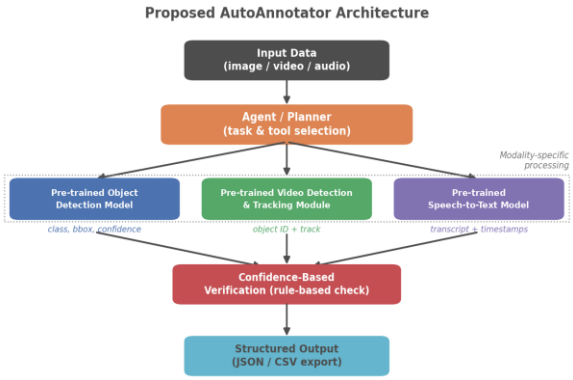

Fig. 3. Proposed AutoAnnotator Architecture.

Input data (image, video, or audio) is routed by the Agent/Planner to the corresponding modality-specific processing component -- a pre-trained object detection model, a pre-trained video detection and tracking module, or a pre-trained speech-to-text model -- after which outputs pass through confidence-based (rule-based) verification before being exported as structured JSON/CSV output. This architecture reflects the design described in the project synopsis; it has not been implemented or evaluated in this paper.

1. *Data input:* the user uploads an image, video or audio file.
2. *Input detection:* the system identifies the data type.
3. *Task selection:* the Agent/Planner determines the annotation task.
4. *Preprocessing:* video frames are extracted and audio is prepared.
5. *Annotation:* a suitable pre-trained model for the corresponding task generates labels, boxes, tracks or transcripts.
6. *Confidence check:* results are checked with confidence values and simple verification rules.
7. *Storage:* validated annotations are stored in a structured format.
8. *Export:* results are displayed and exported as JSON or CSV.
9. *Evaluation:* annotation quality, processing time and manual-effort reduction are planned measures.

Table III. Proposed components (from the synopsis).

Component	Purpose	Output
Pre-trained object detection model	Lightweight object detection	Class, bounding box, confidence
OpenCV	Media processing, frame extraction	Processed media

Component	Purpose	Output
Pre-trained object tracking module	Object tracking	Object ID and track
Pre-trained speech-to-text model	Speech-to-text	Transcript with timestamps
Rule-based confidence check	Identify uncertain results	Accept or flag uncertain results; detailed handling will be defined during implementation.
Agent/Planner	Select task and tool, control workflow	Task plan / tool selection

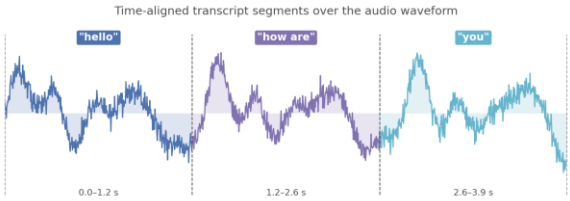

Fig. 4. Conceptual illustration of speech-to-text annotation, showing transcript segments time-aligned to the audio waveform.

The planned implementation uses Python, PyTorch (for loading and running the pre-trained models), OpenCV, NumPy/Pandas and Streamlit or Flask. Google Colab is an optional GPU alternative for running the pre-trained models.

Candidate model families include a pre-trained YOLO (You Only Look Once)-family detector with a multi-object tracker such as ByteTrack for video, and a pre-trained speech recognition model such as Whisper for audio; the final selection will be made during implementation. The synopsis does not specify whether the planner is LLM-driven; this will be defined during implementation.

C. Positioning against the literature

Table IV relates the proposal to the research gaps identified in Section VI.

Table IV. How the proposal relates to the identified gaps (design intent only).

Gap	Proposed AutoAnnotator response	Status
G1	Single workflow for image, video, audio	Proposed
G2	Agent/Planner for task and tool selection	Proposed
G3–G4	Rule-based confidence and consistency checks	Proposed; rules not yet specified
G5	Not in current workflow; active learning from user corrections is future scope	Not proposed for the prototype
G6	Video tracking and speech transcription are in scope	Proposed
G7	Accuracy, time and effort metrics planned	Proposed
G8	Lightweight pre-trained models, optional Colab	Proposed
G9	Common JSON/CSV output	Proposed
G10	Small/medium public datasets and user samples	Proposed

The literature offers several design ideas that are not part of the current project scope. Certainty-based routing to humans [1], multi-model voting [3] and recognition-style question-answering checks [6] are possible verification extensions, while schema-based tool interfaces [8] may inform tool selection. LLM-based planning, multi-agent coordination and retrieval-augmented generation (RAG)-based verification are not specified as features of the current proposed workflow and are not presented here as implemented capabilities.

D. Limitations acknowledged in the synopsis

Accuracy will depend on the chosen pre-trained models and on how well they suit the input data. Complex or domain-specific objects, or classes outside a model's training vocabulary, may be missed; large videos require more time and storage; limited GPU memory may constrain model size and batch processing; and only selected annotation tasks are in scope.

VIII. DISCUSSION

The synthesis indicates that many reviewed systems obtain label information from strong pretrained or foundation models, including CLIP-based transfer [2], [12]–[16] and multi-VLM pipelines [3]. Verification remains a separate challenge: studies add certainty routing, voting or QA checks rather than relying on a single confidence signal. These findings motivate AutoAnnotator's lightweight design, its use of suitable pre-trained models, and its explicit verification layer.

Three cautions apply. First, the corpus does not provide performance evidence for AutoAnnotator's proposed tools or its chosen pre-trained models in this configuration.

Second, agentic AI is treated conceptually through [7], [8], and its effect on annotation efficiency remains untested. Third, the corpus mixes domains and metrics, so cross-paper numeric comparisons are not directly comparable.

Threats to validity. The review covers a bounded corpus of 20 publications and is not exhaustive. Some included papers are adjacent to, rather than directly about, annotation; the 2026 preprints [4]–[6] may change; and the reported findings were not independently reproduced. Accordingly, the research gaps should be interpreted as gaps within this corpus, not as definitive claims about all published work.

IX. FUTURE DIRECTIONS

Future work in the synopsis includes image segmentation, document/text annotation, domain-specific agents, cloud processing, active learning from user corrections, integration with annotation platforms, additional video/audio models, and mobile or cloud versions.

The reviewed literature adds further research directions:

- calibrated confidence and certainty-based human routing [1];
- multi-model voting and recognition-style self-checks as verification layers [3], [6];
- schema-based tool interfaces for the planner [8];
- security and hallucination safeguards for agent components [7];
- shared benchmarks for annotation quality, time and effort across modalities;
- evidence on speech-to-text and video-tracking annotation quality.

X. CONCLUSION

Across the 20 publications in the reviewed corpus, AI methods support individual annotation tasks, while verification strategies remain task-specific and agentic orchestration across modalities is not demonstrated. The literature also highlights the roles of human review, reference comparison and confidence calibration.

In response to these observations, the proposed AutoAnnotator framework combines modality detection, an Agent/Planner, suitable pre-trained models for image, video and audio tasks, confidence-based verification, and JSON/CSV export. The framework remains a proposal: it has not been implemented or evaluated in this paper, so no accuracy, speed or effort-reduction results are claimed. Future work should implement the workflow and evaluate annotation quality, processing time and manual effort on clearly specified datasets.

ACKNOWLEDGMENT

The authors thank the Department of Computer Engineering, Guru Gobind Singh College of Engineering and Research Centre, Nashik, for academic support during this work.

CONFLICT OF INTEREST

The authors declare that they have no conflict of interest.

FUNDING

This research received no external funding.

DECLARATION OF AI USE

An agentic AI tool was used to support language editing, formatting and drafting of this manuscript. The authors reviewed and verified all content, including the literature analysis and references, and take full responsibility for the paper. No AI tool is listed as an author.

REFERENCES

[1] J. Beck, L. M. Kemeter, K. Dürrbeck, M. H. I. Abdalla, and F. Kreuter, "Toward integrating ChatGPT into satellite image annotation workflows: A comparison of label quality and costs of human and automated annotators," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 18, pp. 4366–4381, 2025, doi: 10.1109/JSTARS.2025.3528192.

[2] T. Jain, J. Soni, N. Sumith, and B. Ashwath Rao, "A comparative framework for automatic image annotation using CNN architectures and CLIP-ViT," in Proc. 2026 Int. Conf. Artificial Intelligence and Data Engineering (AIDE), 2026, pp. 497–501, doi: 10.1109/AIDE69088.2026.11544571.

[3] X. Su, H. Chen, K. Liu, Y. Yao, and E. Huang, "Automatic image labeling framework based on multiple vision-language models for automation applications," in Proc. 10th Int. Conf. Robotics and Automation Engineering (ICRAE), 2025, doi: 10.1109/ICRAE67496.2025.00048.

[4] Z. Lahrichi, M. Ferras, G. Richard, and G. Peeters, "SonicCaps: Large-scale diverse and fine-grained captioning for improved audio-retrieval," arXiv:2609.02343, 2026, doi: 10.48550/arXiv.2609.02343.

[5] T. Alex, W. Suharitdamrong, S. Atito, A. Mustafa, M. Awais, P. J. B. Jackson, J. Deng, and I. Elezi, "Listen, reason, and segment: Aligning LALMs with editorial judgment for media chapterization," arXiv:2608.16539, 2026, doi: 10.48550/arXiv.2608.16539.

[6] F. Ma, Y. Rong, X. Li, C. Zhang, P. Wan, and L. Liu, "Closing the verification loop: Self-check captioning for long-paragraph detailed audio captioning," arXiv:2608.30713, 2026, doi: 10.48550/arXiv.2608.30713.

[7] M. A. Ferrag, A. Lakas, N. Tihanyi, and M. Debbah, "LLM and AI agents for autonomous systems: A survey of applications, datasets, and security challenges," IEEE Open J. Intell. Transp. Syst., vol. 7, pp. 615–657, 2026, doi: 10.1109/OJITS.2026.3665677.

[8] T. Ishida, Y. Murakami, D. Lin, and K. M. Lhaksmana, "AI agents: From concept to code to commerce," in Proc. 2025 Int. Conf. Information and Communication Technology (ICoICT), 2025, doi: 10.1109/ICoICT66265.2025.11193130.

[9] A. Satpute, A. Greiner-Petter, M. Schubotz, N. Meuschke, A. Aizawa, O. Teschke, and B. Gipp, "TEIMMA: The first content reuse annotator for text, images, and math," in Proc. 2023 ACM/IEEE Joint Conf. Digital Libraries (JCDL), 2023, doi: 10.1109/JCDL57899.2023.00056.

[10] W. Zimmer, A. Rangesh, and M. Trivedi, "3D BAT: A semi-automatic, web-based 3D annotation toolbox for full-surround, multi-modal data streams," in Proc. 2019 IEEE Intelligent Vehicles Symp. (IV), 2019, pp. 1816–1821, doi: 10.1109/IVS.2019.8814071.

[11] T.-C. Wu, X. Wang, and D. M. Umulis, "Ground truth annotator and 3D dataset generator for validation of nuclei segmentation programs," in Proc. 2019 12th Int. Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), 2019, doi: 10.1109/CISP-BMEI48845.2019.8965881.

[12] S. Peng, K. Genova, C. Jiang, A. Tagliasacchi, M. Pollefeys, and T. Funkhouser, "OpenScene: 3D scene understanding with open vocabularies," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2023, pp. 815–824, doi: 10.1109/CVPR52729.2023.00085.

[13] Y. Lu, C. Xu, X. Wei, X. Xie, M. Tomizuka, K. Keutzer, and S. Zhang, "Open-vocabulary point-cloud object detection without 3D annotation," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2023, doi: 10.1109/CVPR52729.2023.00121.

[14] P.-H. Hsu, K. Zhang, F.-E. Wang, T. Tu, M.-F. Li, Y.-L. Liu, A. Y. C. Chen, M. Sun, and C.-H. Kuo, "OpenM3D: Open vocabulary multi-view indoor 3D object detection without human annotations," in Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV), 2025, pp. 8688–8698, doi: 10.1109/ICCV51701.2025.00813.

[15] L. Chen, X. Wang, J. Lu, S. Lin, C. Wang, and G. He, "CLIP-driven open-vocabulary 3D scene graph generation via cross-modality contrastive learning," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2024, pp. 27863–27873, doi: 10.1109/CVPR52733.2024.02632.

[16] S. Wang, Y. Duan, H. Ding, Y.-P. Tan, K.-H. Yap, and J. Yuan, "Learning transferable human-

object interaction detector with natural language supervision," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2022, pp. 939–948, doi: 10.1109/CVPR52688.2022.00101.

[17] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner, "ScanNet: Richly-annotated 3D reconstructions of indoor scenes," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2432–2443, doi: 10.1109/CVPR.2017.261.

[18] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "PointNet: Deep learning on point sets for 3D classification and segmentation," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2017, pp. 652–660, doi: 10.1109/CVPR.2017.16.

[19] W. Zhou, X. Jiang, and Y.-H. Liu, "MVPointNet: Multi-view network for 3D object based on point cloud," *IEEE Sensors J.*, vol. 19, no. 24, pp. 12145–12152, Dec. 2019, doi: 10.1109/JSEN.2019.2937089.

[20] Z. Zhang and M. Sun, "Research on biomedical image segmentation method based on full convolutional neural network," in *Proc. 2023 IEEE Int. Conf. Sensors, Electronics and Computer Engineering (ICSECE)*, 2023, doi: 10.1109/ICSECE58870.2023.10263584.